

# Optimal Shrinkage of Singular Values

Matan Gavish \*

David L. Donoho \*

## Abstract

We consider recovery of low-rank matrices from noisy data by shrinkage of singular values, in which a single, univariate nonlinearity is applied to each of the empirical singular values. We adopt an asymptotic framework, in which the matrix size is much larger than the rank of the signal matrix to be recovered, and the signal-to-noise ratio of the low-rank piece stays constant. For a variety of loss functions, including the Frobenius norm loss (MSE), nuclear norm loss and operator norm loss, we show that in this framework there is a well-defined asymptotic loss that we evaluate precisely in each case. In fact, each of the loss functions we study admits a *unique admissible* shrinkage nonlinearity dominating all other nonlinearities. We provide a general method for evaluating these optimal nonlinearities, and demonstrate it by working out explicit formulas for the optimal nonlinearities in the Frobenius, nuclear and operator norm cases.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Preliminaries</b>	<b>4</b>
2.1	Natural problem scaling . . . . .	4
2.2	Asymptotic framework and problem statement . . . . .	5
2.3	Our contribution . . . . .	7
<b>3</b>	<b>The Asymptotic Picture</b>	<b>7</b>
<b>4</b>	<b>Optimal Shrinker for Frobenius Loss</b>	<b>9</b>
4.1	Lower bound on asymptotic loss . . . . .	9
4.2	Optimal shrinker matching the lower bound . . . . .	10
<b>5</b>	<b>A Framework for Finding Optimal Shrinkers</b>	<b>11</b>
5.1	Discussion . . . . .	13
5.2	Simultaneous Block Diagonalization . . . . .	13
5.3	Deterministic formula for the asymptotic loss . . . . .	15
<b>6</b>	<b>Optimal Shrinkers for Frobenius, Operator &amp; Nuclear Losses</b>	<b>17</b>
6.1	Frobenius norm loss . . . . .	18
6.2	Operator norm loss . . . . .	19
6.3	Nuclear norm loss . . . . .	19

---

\*Department of Statistics, Stanford University

<b>7</b>	<b>Extensions</b>	<b>20</b>
7.1	Unknown Noise level . . . . .	20
7.2	General white noise . . . . .	21
<b>8</b>	<b>Conclusion</b>	<b>22</b>

# 1 Introduction

Suppose that we are interested in an  $m$  by  $n$  matrix  $X$ , which is thought to be either exactly or approximately of low rank, but we only observe a single noisy  $m$ -by- $n$  matrix  $Y$ , obeying  $Y = X + \sigma Z$ ; The noise matrix  $Z$  has independent, identically distributed entries with zero mean and unit variance. We choose a loss function  $L_{m,n}(\cdot, \cdot)$  and wish to recover the matrix  $X$  with some bound on the risk  $\mathbb{E}L_{m,n}(X, \hat{X})$ , where  $\hat{X}$  is our estimated value of  $X$ .

For example, when choosing the square Frobenius loss

$$L_{m,n}^{fro}(X, \hat{X}) = \left\| X - \hat{X} \right\|_F^2 = \sum_{i,j} |X_{i,j} - \hat{X}_{i,j}|^2, \quad (1)$$

where  $X$  and  $\hat{X}$  are  $m$ -by- $n$  matrices, we would like to find an estimator  $\hat{X}$  with small mean square error (MSE). The default technique for estimating a low rank matrix in noise is the *Truncated SVD* (TSVD) [1]: write

$$Y = \sum_{i=1}^m y_i \mathbf{v}_i \tilde{\mathbf{v}}_i' \quad (2)$$

for the Singular Value Decomposition of the data matrix  $Y$ , where  $\mathbf{v}_i \in \mathbb{R}^m$  and  $\tilde{\mathbf{v}}_i \in \mathbb{R}^n$ ,  $i = 1, \dots, m$  are the left and right singular vectors of  $Y$  corresponding to the singular value  $y_i$ . The TSVD estimator is

$$\hat{X}_r = \sum_{i=1}^r y_i \mathbf{v}_i \tilde{\mathbf{v}}_i',$$

where  $r = \text{rank}(X)$ , assumed known, and  $y_1 \geq \dots \geq y_m$ . Being the best approximation of rank  $r$  to the data in the least squares sense [2], and therefore the Maximum Likelihood estimator when  $Z$  has Gaussian entries, the TSVD is arguably as ubiquitous in science and engineering as linear regression [3–8].

The TSVD estimator shrinks to zero some of the data singular values, while leaving others untouched. More generally, for any specific choice of scalar non-linearity  $\eta : [0, \infty) \rightarrow [0, \infty)$ , also known as a shrinker, there is a corresponding *singular value shrinkage* estimator  $\hat{X}_\eta$  given by

$$\hat{X}_\eta = \sum_{i=1}^m \eta(y_i) \mathbf{v}_i \tilde{\mathbf{v}}_i'. \quad (3)$$

For scalar and vector denoising, univariate shrinkage rules have proved to be simple and practical denoising methods, with near-optimal performance guarantees under various performance measures [9–12]. Shrinkage makes sense for

singular values, too: presumably, the observed singular values  $y_1 \dots y_m$  are “inflated” by the noise, and applying a carefully chosen shrinkage function, one can obtain a good estimate of the original signal  $X$ .

Is there a simple, natural shrinkage nonlinearity for singular values? If there is a simple answer to this question, surely it depends on the loss function  $L$  and on specific assumptions on the signal matrix  $X$ .

In [13] we have performed a narrow investigation that focused on hard and soft thresholding of singular values under the Frobenius loss (1). We adopted a simple asymptotic framework that models the situation where  $X$  is low-rank, originally proposed by Shabalin and Nobel [14] and inspired by Johnstone’s Spiked Covariance Model [15]. In this framework, the signal matrix dimensions  $m = m_n$  and  $n$  both to infinity, such that their ratio converges to an asymptotic aspect ratio:  $m_n/n \rightarrow \beta$ , with  $0 < \beta \leq 1$ , while the column span of the signal matrix remains fixed. Building on a recent probabilistic analysis of this framework [16] we have discovered that, in this framework, there is an *asymptotically unique admissible* threshold for singular values, in the sense that it offers equal or better asymptotic MSE to that of any other threshold, across all possible signal configurations.

The main discovery reported here is that this phenomenon is in fact much more general: in this asymptotic framework, which models low-rank matrices observed in white noise, for each of a variety of loss functions, there exists a single *asymptotically unique admissible* shrinkage nonlinearity, in the sense that it offers equal or better asymptotic loss to that of any other shrinkage nonlinearity, across all possible signal configurations. In other words, once the loss function has been decided, in a definite asymptotic sense, there is a single rational choice of shrinkage nonlinearity.

## Some optimal shrinkers

In this paper, we develop a general method for finding the optimal shrinkage nonlinearity for a variety of loss functions. We explicitly work out the optimal shrinkage formula for the Frobenius norm loss, the nuclear norm loss, and the operator norm loss:

**Frobenius norm loss.** The optimal nonlinearity for the Frobenius norm loss (1) is

$$\eta^*(y) = \begin{cases} \frac{1}{y} \sqrt{(y^2 - \beta - 1)^2 - 4\beta} & y \geq 1 + \sqrt{\beta} \\ 0 & y \leq 1 + \sqrt{\beta} \end{cases}. \quad (4)$$

In the asymptotically square case  $\beta = 1$  this reduces to

$$\eta(y) = \sqrt{(y^2 - 4)_+}.$$

**Operator norm loss** The Operator norm loss is  $L_{m,n}^{op}(X, \hat{X}) = \|X - \hat{X}\|_{op}$ . Here,  $\|A\|_{op}$  is the operator norm of  $A$ , considered as a linear operator  $\mathcal{L}_2(\mathbb{R}^n) \rightarrow \mathcal{L}_2(\mathbb{R}^m)$ , namely, its maximal singular value. The operator norm loss has mostly been studied not in the context of matrix denoising but in the closely related context of covariance estimation [17–19].

Define

$$x(y) = \begin{cases} \frac{1}{\sqrt{2}} \sqrt{y^2 - \beta - 1 + \sqrt{(y^2 - \beta - 1)^2 - 4\beta}} & y \geq 1 + \sqrt{\beta} \\ 0 & y \leq 1 + \sqrt{\beta} \end{cases}. \quad (5)$$

Then the optimal nonlinearity for operator loss is just

$$\eta^*(y) = x(y). \quad (6)$$

**Nuclear norm loss.** The Nuclear norm loss is  $L_{m,n}^{nuc}(X, \hat{X}) = \|X - \hat{X}\|_*$ , where  $\|A\|_*$  is the nuclear norm of the matrix  $A$ , namely the sum of its singular values. See [20] and references within for discussion of the Nuclear norm and more generally of Schatten- $p$  norms as losses for matrix estimation.

The optimal nonlinearity for nuclear norm loss is

$$\eta^*(y) = \begin{cases} \frac{1}{x^2 y} (x^4 - \beta - \sqrt{\beta} xy) & x^4 \geq \beta + \sqrt{\beta} xy \\ 0 & x^4 < \beta + \sqrt{\beta} xy \end{cases}. \quad (7)$$

where  $x = x(y)$  is given in (5).

These nonlinearities are shown in Figure 1. Note that the formulas above are calibrated for the natural noise level  $\sigma = 1/\sqrt{n}$ ; see Section 7.1 below for usage in known noise level  $\sigma$  or unknown noise level.

## 2 Preliminaries

Column vectors are denoted by boldface lowercase letters, such as  $\mathbf{v}$ , their transpose is  $\mathbf{v}'$  and their  $i$ -th coordinate is  $v_i$ . The Euclidean inner product and norm on vectors are denoted by  $\langle \mathbf{u}, \mathbf{v} \rangle$  and  $\|\mathbf{u}\|_2$ , respectively. Matrices are denoted by uppercase letters, such as  $X$ , their transpose is  $X'$  and their  $i, j$ -th entry is  $A_{i,j}$ .  $M_{m \times n}$  denotes the space of real  $m$ -by- $n$  matrices,  $\langle X, Y \rangle = \sum_{i,j} X_{i,j} Y_{i,j}$  denotes the Hilbert-Schmidt inner product and  $\|X\|_F$  denotes the corresponding Frobenius norm on  $M_{m \times n}$ .  $\|X\|_*$  and  $\|X\|_{op}$  denote the nuclear norm (sum of singular values) and operator norm (maximal singular value) of  $X$ , respectively. For simplicity we only consider  $m \leq n$ . We denote matrix denoisers, or estimators, by  $\hat{X} : M_{m \times n} \rightarrow M_{m \times n}$ . The symbols  $\xrightarrow{a.s.}$  and  $\stackrel{a.s.}{=}$  denote almost sure convergence and equality of a.s. limits, respectively. We use “fat” SVD of a matrix  $X \in M_{m \times n}$  with  $m \leq n$ , that is, when writing  $X = U D \tilde{U}$  we mean that  $U \in M_{m \times m}$ ,  $D \in M_{m \times n}$ , and  $\tilde{U} \in M_{n \times n}$ . Symbols without tilde such as  $\mathbf{u}$  are associated with left singular vectors, while symbols with tilde such as  $\tilde{\mathbf{u}}$  are associated with right singular vectors. By  $\text{diag}(x_1, \dots, x_m)$  we mean the  $m$ -by- $n$  matrix whose main diagonal is  $x_1, \dots, x_m$ , with  $n$  implicit in the notation and inferred from context.

### 2.1 Natural problem scaling

In the general model  $Y = X + \sigma Z$ , the noise level in the singular values of  $Y$  is  $\sqrt{n}\sigma$ . Instead of specifying a different shrinkage rule that depends on the matrix size  $n$ , we calibrate our shrinkage rules to the “natural” model  $Y = X + Z/\sqrt{n}$ .

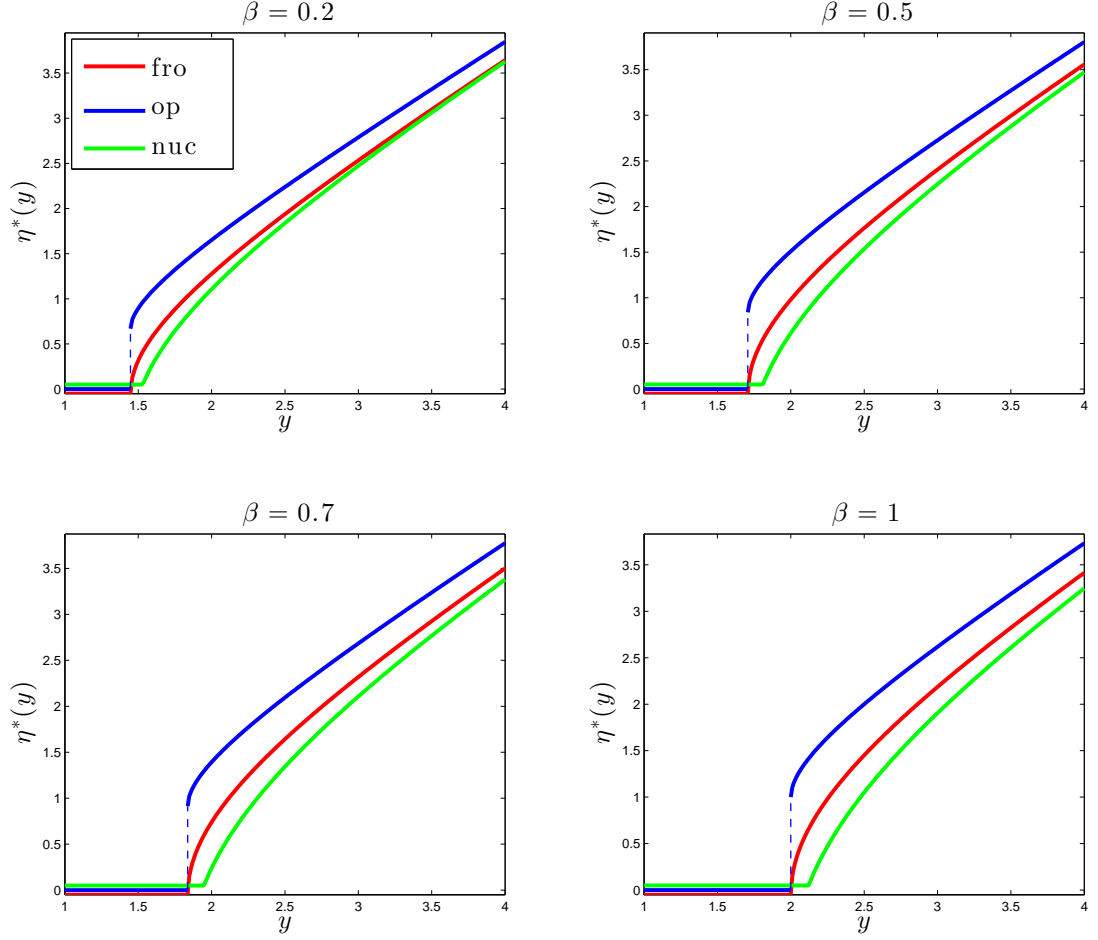


Figure 1: Optimal shrinkers for Frobenius, Operator and Nuclear norm losses for different values of  $\beta$ . All shrinkers are zero on  $y \in [0, 1]$  (not shown) and asymptote to the identity  $\eta(y) = y$  as  $y \rightarrow \infty$ . Curves jittered in the vertical axis to avoid overlap.

In this convention, shrinkage rules stay the same for every value of  $n$ , and we conveniently abuse notation by writing  $\hat{X}_\eta$  as in (3) for any  $\hat{X}_\eta : M_{m \times n} \rightarrow M_{m \times n}$ , keeping  $m$  and  $n$  implicit. To apply any denoiser  $\hat{X}$  below to data from the general model  $Y = X + \sigma Z$ , use the denoiser

$$\hat{X}_\eta^{(n, \sigma)}(Y) = \sqrt{n}\sigma \cdot \hat{X}_\eta(Y/\sqrt{n}\sigma). \quad (8)$$

Throughout the text, we use  $\hat{X}_\eta$  to denote singular value shrinker calibrated for noise level  $1/\sqrt{n}$ . In Section 7.1 below we provide a recipe for applying any denoiser  $\hat{X}_\eta$  calibrated for noise level  $\sigma = 1/\sqrt{n}$  for data in the presence of unknown noise level.

## 2.2 Asymptotic framework and problem statement

In this paper, we consider a sequence of increasingly larger denoising problems

$$Y_n = X_n + Z_n/\sqrt{n} \quad (9)$$

with  $X_n, Z_n \in M_{m_n, n}$ , satisfying the following assumptions:

1. *Invariant white noise*: The entries of  $Z_n$  are i.i.d samples from a distribution with zero mean, unit variance and finite fourth moment. To simplify the formal statement of our results, we assume that this distribution is *orthogonally invariant* in the sense that  $Z_n$  follows the same distribution as  $AZ_nB$ , for every orthogonal  $A \in M_{m_n, m_n}$  and  $B \in M_{n, n}$ . This is the case, for example, when the entries of  $Z_n$  are Gaussian. In Section 7.2 we revisit this restriction and discuss general (not necessarily invariant) white noise.
2. *Fixed signal column span*  $(x_1, \dots, x_r)$ : Let the rank  $r > 0$  be fixed and choose a vector  $\mathbf{x} \in \mathbb{R}^r$  with coordinates  $\mathbf{x} = (x_1, \dots, x_r)$  such that  $x_1 > \dots > x_r > 0$ . Assume that for all  $n$ ,

$$X_n = U_n \text{diag}(x_1, \dots, x_r, 0, \dots, 0) \tilde{U}_n' \quad (10)$$

is an arbitrary singular value decomposition of  $X_n$ , where  $U_n \in M_{m_n, m_n}$  and  $\tilde{U}_n \in M_{n, n}$ .

3. *Asymptotic aspect ratio*  $\beta$ : The sequence  $m_n$  is such that  $m_n/n \rightarrow \beta$ . To simplify our formulas, we assume that  $0 < \beta \leq 1$ .

#### Remarks.

- While the signal rank  $r$  and nonzero signal singular values  $x_1, \dots, x_r$  are shared by all matrices  $X_n$ , the signal left and right singular vectors  $U_n$  and  $V_n$  are unknown and arbitrary.
- The assumption that the signal singular values are non-degenerate ( $x_i > x_{i+1}$ ,  $1 \leq i < r$ ) is not necessary for our results to hold, yet it simplifies the analysis considerably.

**Definition 1. Asymptotic Loss.** Let  $L = \{L_{m,n} \mid (m, n) \in \mathbb{N} \times \mathbb{N}\}$  be a family of losses, where each  $L_{m,n} : M_{m \times n} \times M_{m \times n} \rightarrow [0, \infty)$  is a loss function obeying  $L_{m,n}(X, X) = 0$ . Let  $\eta$  be a continuous nonlinearity and consider  $\hat{X}_\eta$ , the singular value shrinkage denoiser (3) calibrated, as discussed above, for noise level  $1/\sqrt{n}$ . Let  $m_n$  be an increasing sequence such that  $\lim_{n \rightarrow \infty} m_n/n = \beta$ , implicit in our notation. Define the *asymptotic loss* of the shrinker  $\eta$  (with respect to  $L$ ) at the signal  $\mathbf{x} = (x_1, \dots, x_r)$  by

$$L_\infty(\eta|\mathbf{x}) \stackrel{\text{a.s.}}{=} \lim_{n \rightarrow \infty} L_{m_n, n} \left( X_n, \hat{X}_\eta(X_n + \frac{1}{\sqrt{n}} Z_n) \right).$$

Our results imply that the AMSE is well-defined as a function of the signal singular values  $\mathbf{x}$ .

**Definition 2. Optimal Shrinker.** If a continuous shrinker  $\eta^*$  satisfies

$$L_\infty(\eta^*|\mathbf{x}) \leq L_\infty(\eta|\mathbf{x})$$

for any other continuous shrinker  $\eta$ , any  $r \geq 1$  and any  $\mathbf{x} \in \mathbb{R}^r$ , we say that  $\eta^*$  is *unique asymptotically admissible* (of simply “optimal”) for the loss sequence  $L$ .

### 2.3 Our contribution

At first glance, it seems too much to hope that optimal shrinkers in the sense of Definition 2 even exist. Indeed, existence of an optimal shrinker for a loss family  $L$  implies that, asymptotically, the decision-theoretic picture is extremely simple and actionable: from the asymptotic loss perspective, there is a single rational choice for shrinker.

In our current terminology, Shabalin and Nobel [14] have effectively conjectured that an optimal shrinker exists for Frobenius loss and provided a proof outline. The estimator they derive can be shown to be equivalent to the optimal shrinker (4), yet was given in a more complicated form. (In Section 4 we visit the special case of Frobenius loss in detail, and prove that (4) is the optimal shrinker.)

Our contribution in this paper is as follows.

1. We rigorously establish the existence of an optimal shrinker for a variety of loss families, including the popular Frobenius, operator and nuclear norm losses.
2. We provide a framework for finding the optimal shrinkers for a variety of loss families including these popular losses. As discussed in Section 7.1, our framework can be applied whether the noise level  $\sigma$  is known or unknown.
3. We use our framework to find simple, explicit formulas for the optimal shrinkers for Frobenius, operator and nuclear norm losses.

In the related problem of covariance estimation in the Spiked Covariance Model, in collaboration with I. Johnstone we identified a similar phenomenon, namely, existence of optimal shrinkage for eigenvalues for the purpose of covariance estimation [21].

## 3 The Asymptotic Picture

In the “null case”  $X_n \equiv 0$ , the empirical distribution of the singular values of  $Y_n = Z_n/\sqrt{n}$  famously converge, as  $n \rightarrow \infty$ , to the quarter-circle distribution if  $\beta = 1$ , or to the generalized quarter-circle if  $\beta < 1$  [22]. This distribution is compactly supported on  $[\beta_-, \beta_+]$ , with

$$\beta_{\pm} = 1 \pm \sqrt{\beta}.$$

We say that the singular values of  $Y_n$  forms a (generalized) quarter circle *bulk* and call  $\beta_+$  the *bulk edge*.

Expanding seminal results of [23, 24] and others, recently Benaych-Georges and Nadakuditi [16] have provided a thorough analysis of a collection of models, which includes the model (9) as a special case. In this section we summarize some of their results regarding asymptotic behaviour of the model (9), which are relevant to singular value shrinkage.



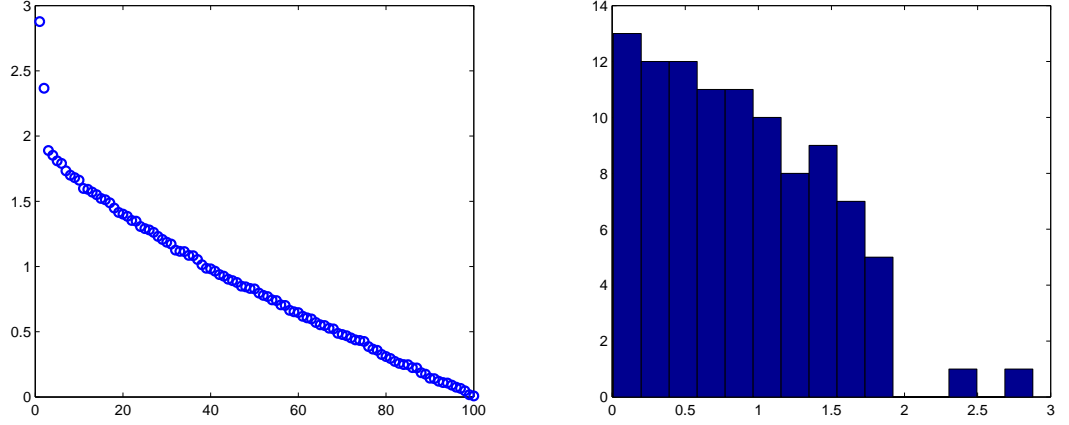


Figure 2: Singular values of a data matrix  $Y \in M_{100,100}$  drawn from the model  $Y = X + Z/\sqrt{100}$ , with  $r = 2$  and  $\mathbf{x} = (2.5, 1.7)$ . Left: singular values in decreasing order. Right: Histogram of the singular values (note the bulk edge close to 2).

For  $x \geq \beta^{1/4}$ , define

$$y(x) = \sqrt{\left(x + \frac{1}{x}\right) \left(x + \frac{\beta}{x}\right)}, \quad (11)$$

$$c(x) = \sqrt{\frac{x^4 - \beta}{x^4 + \beta x^2}} \quad \text{and} \quad (12)$$

$$\tilde{c}(x) = \sqrt{\frac{x^4 - \beta}{x^4 + x^2}}. \quad (13)$$

It turns out that  $y(x)$  from Eq. (11) is the asymptotic location of a data singular value corresponding to a signal singular value  $x$ , provided  $x \geq \beta^{1/4}$ . (Note that the function  $x(y)$  from (5) is the inverse of  $y(x)$  when  $x \geq \beta^{1/4}$ , and that  $y(\beta^{1/4}) = \beta_+$ .) Similarly,  $c(x)$  from Eq. (12) (resp.  $\tilde{c}(x)$  from Eq. (13)) is the cosine of the asymptotic angle between the signal left (resp. right) singular vector and the corresponding data left (resp. right) singular vector, provided that the corresponding signal singular value  $x$  satisfies  $x \geq \beta^{1/4}$ .

Additional notation is required to state these facts formally. We rewrite the sequence of signal matrices in our asymptotic framework (10) as

$$X_n = \sum_{i=1}^r x_i \mathbf{u}_{n,i} \tilde{\mathbf{u}}'_{n,i}, \quad (14)$$

so that  $\mathbf{u}_{n,i} \in \mathbb{R}^{m_n}$  (resp.  $\tilde{\mathbf{u}}_{n,i} \in \mathbb{R}^n$ ) is the left (resp. right) singular vector corresponding to the singular value  $x_i$ , namely,  $i$ -th column of  $U_n$  (resp.  $\tilde{U}_n$ ) in (10). Similarly, let  $Y_n$  be a corresponding sequence of observed matrices in our framework, and write

$$Y_n = \sum_{i=1}^{m_n} y_{n,i} \mathbf{v}_{n,i} \tilde{\mathbf{v}}'_{n,i} \quad (15)$$

so that  $\mathbf{v}_{n,i} \in \mathbb{R}^m$  (resp.  $\tilde{\mathbf{v}}_{n,i} \in \mathbb{R}^n$ ) is the left (resp. right) singular vector corresponding to the singular value  $y_{n,i}$ .

In this notation, [16] have proved:



**Lemma 1. Asymptotic location of the top  $r$  data singular values.** For  $1 \leq i \leq r$ ,

$$\lim_{n \rightarrow \infty} y_{n,i} \stackrel{a.s.}{=} \begin{cases} y(x_i) & x_i \geq \beta^{1/4} \\ \beta_+ & x_i < \beta^{1/4} \end{cases}. \quad (16)$$

It should be noted that the  $m_n - r$  lower singular values of  $Y_n$ , converge almost surely in distribution to the same generalized quarter-circle distribution as in the null case (Figure 2), and moreover that

$$\lim_{n \rightarrow \infty} y_{n,r+1} \stackrel{a.s.}{=} \beta_+. \quad (17)$$

**Lemma 2. Asymptotic angle between signal and data singular vectors.** Let  $1 \leq i \neq j \leq r$  and assume that  $x_i \geq \beta^{1/4}$  is nondegenerate, namely, the value  $x_i$  appears only once in  $\mathbf{x}$ . Then

$$\lim_{n \rightarrow \infty} |\langle \mathbf{u}_{n,i}, \mathbf{v}_{n,j} \rangle| \stackrel{a.s.}{=} \begin{cases} c(x_i) & x_i = x_j \\ 0 & x_i \neq x_j \end{cases}, \quad (18)$$

and

$$\lim_{n \rightarrow \infty} |\langle \tilde{\mathbf{u}}_{n,i}, \tilde{\mathbf{v}}_{n,j} \rangle| \stackrel{a.s.}{=} \begin{cases} \tilde{c}(x_i) & x_i = x_j \\ 0 & x_i \neq x_j \end{cases}. \quad (19)$$

If however  $x_i < \beta^{1/4}$ , then we have

$$\lim_{n \rightarrow \infty} |\langle \mathbf{u}_{n,i}, \mathbf{v}_{n,j} \rangle| \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} |\langle \tilde{\mathbf{u}}_{n,i}, \tilde{\mathbf{v}}_{n,j} \rangle| \stackrel{a.s.}{=} 0.$$

## 4 Optimal Shrinker for Frobenius Loss

As an introduction to the more general framework developed below, we first examine the Frobenius loss case, following the work of Shabalin and Nobel [14].

### 4.1 Lower bound on asymptotic loss

Directly expanding the Frobenius matrix norm, we obtain:

**Lemma 3. Frobenius loss of singular value shrinkage.** For any shrinker  $\eta : [0, \infty) \rightarrow [0, \infty)$ , we have

$$\left\| X_n - \hat{X}_\eta(X_n + Z_n/\sqrt{n}) \right\|_F^2 = \sum_{i=1}^r [x_i^2 + (\eta(y_{n,i}))^2] \quad (20)$$

$$- 2 \sum_{i,j=1}^r x_i \eta(y_{n,i}) \langle \mathbf{u}_{n,i}, \mathbf{v}_{n,j} \rangle \langle \tilde{\mathbf{u}}_{n,i}, \tilde{\mathbf{v}}_{n,j} \rangle \quad (21)$$

$$+ \sum_{i=r+1}^{m_n} (\eta(y_{n,i}))^2 \quad (22)$$

This implies a lower bound on Frobenius loss of any singular value shrinker:

**Correlary 1.** For any shrinker  $\eta : [0, \infty) \rightarrow [0, \infty)$ , we have

$$\left\| X_n - \hat{X}_\eta(X_n + Z_n/\sqrt{n}) \right\|_F^2 \geq \sum_{i=1}^r [x_i^2 + (\eta(y_{n,i}))^2] - 2 \sum_{i,j=1}^r x_i \eta(y_{n,i}) \langle \mathbf{u}_{n,i}, \mathbf{v}_{n,j} \rangle \langle \tilde{\mathbf{u}}_{n,i}, \tilde{\mathbf{v}}_{n,j} \rangle.$$

As  $n \rightarrow \infty$ , this lower bound on the Frobenius loss is governed by three quantities: the asymptotic location of data singular value  $y_{n,i}$ , the asymptotic angle between the left signal singular vectors and left data singular vector  $\langle \mathbf{u}_{n,i}, \mathbf{a}_{n,i} \rangle$ , and asymptotic angle between the right signal singular vectors and right data singular vector  $\langle \mathbf{v}_{n,i}, \mathbf{b}_{n,i} \rangle$  (see also [13]).

Combining Corollary 1, Lemma 1 and Lemma 2 we obtain a lower bound for the asymptotic Frobenius loss:

**Correlary 2.** For any continuous shrinker  $\eta : [0, \infty) \rightarrow [0, \infty)$ , we have

$$L_\infty(\eta|\mathbf{x}) \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} \left\| X_n - \hat{X}_\eta(X_n + Z_n/\sqrt{n}) \right\|_F^2 \geq \sum_{i=1}^r f(\eta|x_i)$$

where

$$f(\eta|x) = x^2 + \eta^2 - 2x\eta c(x)\tilde{c}(x).$$

and  $\eta = \eta(y)$ .

## 4.2 Optimal shrinker matching the lower bound

By differentiating the asymptotic lower bound w.r.t  $\eta$ , we find that  $L_\infty(\eta|\mathbf{x}) \geq \sum_{i=1}^r f(\eta^*|x_i)$ , where  $\eta^*(y(x)) = xc(x)\tilde{c}(x)$ . Expanding  $c(x)$  and  $\tilde{c}(x)$  from Eqs. (12) and (13), we find that  $\eta^*(y)$  is given by (4).

The singular value shrinker  $\eta^*$ , for which  $\hat{X}_{\eta^*}$  minimizes the asymptotic lower bound, thus becomes a natural candidate for the optimal shrinker for Frobenius loss. Indeed, by definition, for  $\hat{X}_{\eta^*}$  the limits of (20) and (21) are the smallest possible. It remains to show that the limit of (22) is the smallest possible.

It is clear from (22) that a necessary condition for a shrinker  $\eta$  to be successful, let alone optimal, is that it must set to zero data eigenvalues that do not correspond to signal. With (17) in mind, we should only consider shrinkers  $\eta$  for which  $\eta(y) = 0$  for any  $y \leq \beta_+$ :

**Definition 3.** A shrinker  $\eta : [0, \infty) \rightarrow [0, \infty)$  is said to *collapse the bulk to 0* if  $\eta(y) = 0$  whenever  $y \leq \beta_+$ .

The following lemma gives a sufficient condition for a shrinker to achieve the lowest limit possible in the term (22), namely, for this term to converge to zero.

**Definition 4.** Assume that a continuous shrinker  $\eta : [0, \infty) \rightarrow [0, \infty)$  collapses the bulk to 0 and satisfies  $\eta(y) \leq C \cdot y$  for some  $C > 0$  and all  $y \geq 0$ . We say that  $\eta$  is a *proper shrinker*.

**Lemma 4.** Let  $\eta : [0, \infty) \rightarrow [0, \infty)$  be a proper shrinker. Then

$$\sum_{i=r+1}^{m_n} (\eta(y_{n,i}))^2 \stackrel{a.s.}{\rightarrow} 0.$$

*Proof.* Assume first that  $r = 1$ . Define  $\beta_+ = 1 + \sqrt{\beta}$ . We may assume that  $X = \text{diag}(1, 0, \dots, 0)$ . Let  $\Pi_n : \mathbb{R}^{m_n} \rightarrow \mathbb{R}^{m_n-1}$  denote the projection on the last  $m_n - 1$  coordinates. Let  $\bar{Y}_n = \Pi_n Y_n = \Pi_n Z_n / \sqrt{n}$  and denote its singular values by  $\bar{y}_{n,1}, \dots, \bar{y}_{n,m_n-1}$ . In other words,  $y_{n,1}^2, \dots, y_{n,m_n}^2$  are the eigenvalues of  $Y_n Y_n'$  and  $\bar{y}_{n,1}^2, \dots, \bar{y}_{n,m_n-1}^2$  are the eigenvalues of

$$\Pi_n Y_n Y_n' \Pi_n' = (\Pi_n Y_n)(\Pi_n Y_n)'.$$

By the Cauchy Interlacing Theorem [25, Theorem 8.1.7], we have  $y_{n,i}^2 \leq \bar{y}_{n,i-1}^2$ ,  $i = 2, \dots, m_n$ . It is therefore enough to show

$$\sum_{i=1}^{m_n-1} (\eta(\bar{y}_{n,i}))^2 \xrightarrow{a.s.} 0.$$

Since by assumption  $\eta$  collapses the bulk to 0, we have

$$\sum_{i=1}^{m_n-1} (\eta(\bar{y}_{n,i}))^2 = \sum_{i=1}^{m_n-1} (\eta(\bar{y}_{n,i}))^2 \cdot \mathbf{1}_{[\beta_+, \infty)}(\bar{y}_{n,i}) \leq C^2 \sum_{i=1}^{m_n-1} (\bar{y}_{n,i})^2 \cdot \mathbf{1}_{[\beta_+, \infty)}(\bar{y}_{n,i})$$

where  $C$  is such that  $\eta(y) \leq Cy$ ,  $y \geq 0$ . However, as almost surely the empirical distribution of  $\{\bar{y}_{n,i}^2\}$  converges in distribution to the Marcenko-Pastur density, we have

$$\lim_{n \rightarrow \infty} \mathbb{E} \sum_{i=1}^{m_n-1} (\bar{y}_{n,i})^2 \cdot \mathbf{1}_{[\beta_+, \infty)}(\bar{y}_{n,i}) = 0.$$

□

The candidate  $\eta^*$  is clearly a proper shrinker in the sense of Definition 4. Lemma 4 thus implies that  $\eta^*$  is the optimal shrinker for Frobenius loss.

## 5 A Framework for Finding Optimal Shrinkers

Our main result may be summarized as follows: the basic ingredients that enabled us to find the optimal shrinker for Frobenius loss allow us to find the optimal shrinker for each of a variety of loss families. For these loss families, an optimal (proper) shrinker exists and is given by a simple formula.

To get started, let us describe the loss families to which our method applies.

**Definition 5. Orthogonally invariant loss.** A loss  $L_{m,n}(\cdot, \cdot)$  is *orthogonally invariant* if for all  $m, n$  we have  $L_{m,n}(A, B) = L_{m,n}(UAV, UBV)$ , for any orthogonal  $U \in O_m$  and  $V \in O_n$ .

**Definition 6. Regular loss.** A loss  $L_{m,n}(\cdot, \cdot)$  is *regular* if for some  $\alpha > 0$ , all  $m, n$  and all  $x > 0$ ,  $L_{m,n}^\alpha(\cdot, \cdot)$  is Lipschitz continuous, in each argument, on the set  $\{(A, B) \mid \sigma_{\min}(A), \sigma_{\min}(B) \geq x\}$ , with Lipschitz constant that may depend on  $x$ . (Here,  $\lambda_{\min}(A)$  is the smallest singular value of  $A$ .)

**Definition 7. Decomposable loss family.** Let  $A, B \in M_{m \times n}$  and let  $m = \sum_{i=1}^k m_i$  and  $n = \sum_{i=1}^k n_i$ . Assume that there are matrices  $A_i, B_i \in M_{m_i, n_i}$ ,  $1 \leq i \leq k$ , such that

$$A = \oplus_i A_i \quad B = \oplus_i B_i$$

in the sense that  $A$  and  $B$  are block-diagonal with blocks  $\{A_i\}$  and  $\{B_i\}$ , respectively. A loss family  $L = \{L_{m,n}\}$  is *sum-decomposable* if, for all  $m, n$  and  $A, B$  with block diagonal structure as above,

$$L_{m,n}(A, B) = \sum_i L_{m_i, n_i}(A_i, B_i).$$

Similarly, it is *max-decomposable* if

$$L_{m,n}(A, B) = \max_i L_{m_i, n_i}(A_i, B_i).$$

**Examples.** As primary examples, we consider loss families defined in Section 1: The Frobenius norm loss  $L^{fro}$ , the operator norm loss  $L^{op}$  and the nuclear norm loss  $L^{nuc}$ . It is easy to check that: (i) any of these losses are orthogonally invariant and regular, and (ii) that the families  $L^{fro}$  and  $L^{nuc}$  are sum-decomposable, while the family  $L^{op}$  is max-decomposable.

Our framework for finding optimal shrinkers can now be stated as follows.

**Theorem 1. Characterization of the optimal singular value shrinker.** *Let*

$$A(x) = \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \quad (23)$$

$$B(\eta, x) = \eta \begin{bmatrix} c(x) \tilde{c}(x) & c(x) \tilde{s}(x) \\ \tilde{c}(x) s(x) & s(x) \tilde{s}(x) \end{bmatrix}, \quad (24)$$

where  $c(x)$  and  $\tilde{c}(x)$  are given by Eqs. (12) and (13), and where  $s(x) = \sqrt{1 - c^2(x)}$  and  $\tilde{s}(x) = \sqrt{1 - \tilde{c}^2(x)}$ . Assume that  $L = \{L_{m,n}\}$  is a sum- or max- decomposable family of regular and orthogonally invariant losses. Define

$$F(\eta, x) = L_{2,2}(A(x), B(\eta, x)) \quad (25)$$

and suppose that for any  $x \geq \beta^{1/4}$  there exists a unique minimizer

$$\eta^{**}(x) = \operatorname{argmin}_{\eta \geq 0} F(\eta, x), \quad (26)$$

such that  $\eta^{**}$  is a proper shrinker on  $[\beta^{1/4}, \infty)$ . Further suppose that there exists a point  $x_0 \geq \beta^{1/4}$  such that

$$F(\eta^{**}(x), x) \geq L_{1,1}(x, 0) \quad \beta^{1/4} \leq x \leq x_0,$$

with

$$F(\eta^{**}(x_0), x_0) = L_{1,1}(x_0, 0). \quad (27)$$

Define the shrinker

$$\eta^*(y) = \begin{cases} \eta^{**}(x(y)) & y(x_0) \leq y \\ 0 & 0 \leq y < y(x_0) \end{cases}, \quad (28)$$

where  $x(y)$  is defined in Eq. (5). Then for any proper shrinker  $\eta$ , the asymptotic losses  $L_\infty(\eta^*|\cdot)$  and  $L_\infty(\eta|\cdot)$  exist, and

$$L_\infty(\eta^*|\mathbf{x}) \leq L_\infty(\eta|\mathbf{x})$$

for all  $r \geq 1$  and all  $\mathbf{x} \in \mathbb{R}^r$ .

## 5.1 Discussion

Before we proceed to prove Theorem 1, we review the information it encodes about the problem at hand and its operational meaning. Theorem 1 is based on a few simple observations:

- First, if  $L$  is a sum- (resp. max-) decomposable family of regular and orthogonally invariant losses, and if  $\eta$  is a proper shrinker, then the asymptotic loss  $L_\infty(\eta|\mathbf{x})$  at  $\mathbf{x} = (x_1, \dots, x_r)$  can be written as a sum (resp. a maximum) over  $r$  terms. These terms have identical functional form. When  $x_i \geq \beta^{1/4}$ , these terms have the form  $L_{2,2}(A(x_i), B(\eta, x_i))$ , and when  $0 \leq x_i < \beta^{1/4}$ , these terms have the form  $L_{1,1}(x_i, 0) + L_{1,1}(0, \eta)$  (resp.  $\max\{L_{1,1}(x_i, 0), L_{1,1}(0, \eta)\}$ ). As a result, one finds that the zero shrinker  $\eta \equiv 0$  is necessarily optimal for  $0 \leq x \leq \beta^{1/4}$ . For  $x \geq \beta^{1/4}$ , one just needs to minimize the loss of a specific 2-by-2 matrix, namely the function  $F$  from (25), to obtain the shrinker  $\eta^{**}$  of (26).
- Second, the asymptotic loss curve  $L_\infty(\eta^{**}|x)$  necessarily crosses the asymptotic loss curve of the zero shrinker  $L_\infty(\eta \equiv 0|x)$  at a point  $x_0 \geq \beta^{1/4}$ .
- Finally, by concatenating the zero shrinker and the shrinker  $\eta^{**}$  precisely at the point  $x_0$  where their asymptotic losses cross, one obtains a shrinker which is continuous ( $x_0 > \beta^{1/4}$ ) or possibly discontinuous ( $x_0 = \beta^{1/4}$ ). However, this shrinker always has a well-defined asymptotic loss. This loss dominates the asymptotic loss of any proper shrinker.

For some loss families  $L = \{L_{m,n}\}$ , it is possible to find an explicit formula for the optimal shrinker using the following steps:

1. Write down an explicit expression for the function  $F(\eta, x)$  from (25).
2. Explicitly solve for the minimizer  $\eta^{**}(x)$  from (26).
3. Write down an explicit expression for the minimum  $F(\eta^{**}(x), x)$
4. Solve (27) for the crossing point  $x_0$ .
5. Compose  $\eta^{**}(x)$  with the transformation  $x(y)$  from (5) to obtain an explicit form of the optimal shrinker  $\eta^*y$  from (28).

In Section 6 we offer three examples of this process. Note that even when an explicit analytic formula is not found, these steps are easily implemented numerically. This allows one to find the value of the optimal shrinker  $\eta^*(y)$  at a desired value  $y$  to any desired numerical precision.

In the remainder of this section we describe a sequence of constructions and lemmas leading to the proof of Theorem 1.

## 5.2 Simultaneous Block Diagonalization

Let us start by considering a fixed signal matrix and noise matrix, without placing them in a sequence. To allow a gentle exposition of the main ideas, we initially make two simplifying assumptions: first, that  $r = 1$ , namely that  $X$  is rank-1, and second, that  $\eta$  shrinks to zero all but the first singular values of  $Y$ , namely,

$\eta(y_i) = 0, i > 1$ . Let  $X \in M_{m \times n}$  be a signal matrix and let  $Y = X + Z/\sqrt{n} \in M_{m \times n}$  be a corresponding data matrix. Denote their SVD by

$$X = U \cdot \text{diag}(x_1, 0, \dots, 0) \cdot \tilde{U} \quad Y = V \cdot \text{diag}(y_1, \dots, y_m) \cdot \tilde{V}.$$

Write  $0_{m \times n}$  for the  $m$ -by- $n$  matrix whose entries are all zeros. The basis pairs  $U, \tilde{U}$  and  $V, \tilde{V}$  diagonalize  $X$  and  $Y$  in the sense that

$$U' X \tilde{U} = x_1 \mathbf{e}_1 \mathbf{e}_1' \oplus 0_{m-1 \times n-1} \quad V' \hat{X}_\eta(Y) \tilde{V} = \eta(y_1) \mathbf{e}_1 \mathbf{e}_1' \oplus 0_{m-1 \times n-1}$$

where  $\mathbf{e}_i$  is  $i$ -th standard basis vector in the appropriate dimension.

Combining these two pairs, we are lead to the following “common” basis pair, which we will denote by  $W, \tilde{W}$ : Let  $w_1, \dots, w_m$  denote the orthonormal basis constructed by applying the GramSchmidt process to the sequence  $\mathbf{u}_1, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{m-1}$ , where  $\mathbf{u}_i$  is the  $i$ -th column of  $U$ , namely the  $i$ -th left singular vector of  $X$ , and similarly,  $\mathbf{v}_i$  is the  $i$ -th column of  $V$ . Denote by  $W$  the matrix whose columns are  $\mathbf{w}_1, \dots, \mathbf{w}_m$ . Repeating this construction for  $\tilde{U}$  and  $\tilde{V}$ , let  $\tilde{w}_1, \dots, \tilde{w}_n$  denote the orthonormal basis constructed by applying the Gram-Schmidt process to the sequence  $\tilde{\mathbf{u}}_1, \tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \dots, \tilde{\mathbf{v}}_{m-1}$ , where  $\tilde{\mathbf{u}}_i$  is the  $i$ -th column of  $U$ , namely the  $i$ -th right singular vector of  $X$ , and similarly,  $\tilde{\mathbf{v}}_i$  is the  $i$ -th column of  $\tilde{V}$ . Denote by  $\tilde{W}$  the matrix whose columns are  $\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_n$ .

We may assume that  $\mathbf{v}_2$  lies in the linear span of  $\mathbf{u}_1$  and  $\mathbf{v}_1$ , and choose the orientation of  $\mathbf{w}_2$  so that

$$\begin{aligned} \mathbf{u}_1 = \mathbf{w}_1 &= c_1 \mathbf{v}_1 + s_1 \mathbf{v}_2 \\ \mathbf{w}_2 &= s_1 \mathbf{v}_1 - c_1 \mathbf{v}_2, \end{aligned}$$

where  $c_1 = \langle \mathbf{u}_1, \mathbf{v}_1 \rangle$  and  $s_1 = \sqrt{1 - c_1^2}$ . Similarly, we may choose the orientation of  $\tilde{\mathbf{w}}_2$  such that

$$\begin{aligned} \tilde{\mathbf{u}}_1 = \tilde{\mathbf{w}}_1 &= \tilde{c}_1 \tilde{\mathbf{v}}_1 + \tilde{s}_1 \tilde{\mathbf{v}}_2 \\ \tilde{\mathbf{w}}_2 &= \tilde{s}_1 \tilde{\mathbf{v}}_1 - \tilde{c}_1 \tilde{\mathbf{v}}_2, \end{aligned}$$

where  $\tilde{c} = \langle \tilde{\mathbf{u}}_1, \tilde{\mathbf{v}}_1 \rangle$  and  $\tilde{s} = \sqrt{1 - \tilde{c}^2}$ . Writing now  $X$  and  $\hat{X}_\eta$  in our new basis pair  $W, \tilde{W}$  we get

$$W' X \tilde{W} = \begin{bmatrix} x_1 & 0 \\ 0 & 0 \end{bmatrix} \oplus 0_{m-2 \times n-2} \quad (29)$$

$$W' \hat{X}_\eta(Y) \tilde{W} = \eta(y_1) \begin{bmatrix} c_1 \tilde{c}_1 & c_1 \tilde{s}_1 \\ \tilde{c}_1 s_1 & s_1 \tilde{s}_1 \end{bmatrix} \oplus 0_{m-2 \times n-2}. \quad (30)$$

It is convenient to rewrite this as

$$W' X \tilde{W} = A(x_1) \oplus 0_{m-2 \times n-2} \quad (31)$$

$$W' \hat{X}_\eta(Y) \tilde{W} = B(\eta(y_1), c_1, s_1, \tilde{c}_1, \tilde{s}_1) \oplus 0_{m-2 \times n-2} \quad (32)$$

where

$$A(x) = \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \quad (33)$$

$$B(\eta, c, s, \tilde{c}, \tilde{s}) = \eta \begin{bmatrix} c\tilde{c} & c\tilde{s} \\ \tilde{c}s & s\tilde{s} \end{bmatrix} \quad (34)$$

Thus, if  $L = \{L_{m,n}\}$  is a sum- or max-decomposable family of orthogonally invariant functions, we have

$$\begin{aligned} L_{m,n}(X, \hat{X}_\eta(Y)) &= L_{m,n}(W'X\tilde{W}, W'\hat{X}_\eta(Y)\tilde{W}) \\ &= L_{m,n}\left(A(x_1) \oplus 0_{m-2 \times n-2}, B(\eta(y_1), c_1, s_1, \tilde{c}_1, \tilde{s}_1) \oplus 0_{m-2 \times n-2}\right) \\ &= L_{2,2}\left(A(x_1), B(\eta(y_1), c_1, s_1, \tilde{c}_1, \tilde{s}_1)\right). \end{aligned}$$

We have proved:

**Lemma 5.** Let  $X = x_1 \mathbf{u}_1 \tilde{\mathbf{u}}_1' \in M_{m \times n}$  be rank-1 and assume that  $Y = \sum_{i=1}^m y_i \mathbf{v}_i \tilde{\mathbf{v}}_i'$  and  $\eta$  are such that  $\eta(y_i) = 0, i > 1$ , where  $y_i$  is the  $i$ -th largest singular value of  $Y$ . Let  $L = \{L_{m,n}\}$  be a sum- or max-decomposable family of orthogonally invariant functions. Then

$$L_{m,n}(X, \hat{X}_\eta(Y)) = L_{2,2}\left(A(x_1), B(\eta(y_1), c_1, s_1, \tilde{c}_1, \tilde{s}_1)\right),$$

where

$$\begin{aligned} c_1 &= \langle \mathbf{u}_1, \mathbf{v}_1 \rangle & s_1 &= \sqrt{1 - c_1^2} \\ \tilde{c}_1 &= \langle \tilde{\mathbf{u}}_1, \tilde{\mathbf{v}}_1 \rangle & \tilde{s}_1 &= \sqrt{1 - \tilde{c}_1^2}. \end{aligned}$$

A similar yet significantly more technical proof gives a similar statement for rank- $r$  matrix  $X$  with non-degenerate singular values (see [21]):

**Lemma 6. A decomposition for the loss.** Let  $X = \sum_{i=1}^r x_i \mathbf{u}_i \tilde{\mathbf{u}}_i' \in M_{m \times n}$  be rank- $r$  with  $x_1 > \dots > x_r > 0$ , and assume that  $Y = \sum_{i=1}^m y_i \mathbf{v}_i \tilde{\mathbf{v}}_i'$  and  $\eta$  are such that  $\eta(y_i) = 0, i > r$ , where  $y_i$  is the  $i$ -th largest singular value of  $Y$ . Let  $L = \{L_{m,n}\}$  be a sum- or max-decomposable family of orthogonally invariant functions. Then

$$L_{m,n}(X, \hat{X}_\eta(Y)) = \sum_{i=1}^r L_{2,2}\left(A(x_i), B(\eta(y_i), c_i, s_i, \tilde{c}_i, \tilde{s}_i)\right),$$

if  $L$  is sum-decomposable, or

$$L_{m,n}(X, \hat{X}_\eta(Y)) = \max_{i=1}^r L_{2,2}\left(A(x_i), B(\eta(y_i), c_i, s_i, \tilde{c}_i, \tilde{s}_i)\right),$$

if  $L$  is max-decomposable. Here,

$$\begin{aligned} c_i &= \langle \mathbf{u}_i, \mathbf{v}_i \rangle & s_i &= \sqrt{1 - c_i^2} \\ \tilde{c}_i &= \langle \tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i \rangle & \tilde{s}_i &= \sqrt{1 - \tilde{c}_i^2}, \end{aligned}$$

for  $i = 1, \dots, r$ .

### 5.3 Deterministic formula for the asymptotic loss

In Section 5.2 we analyzed a single matrix and shown that, for fixed  $m$  and  $n$ , the loss  $L_{m,n}(X, \hat{X}_\eta(Y))$  decomposes to “atomic” units of the form

$$L_{2,2}(A(x_i), B(\eta(y_i), c_i, s_i, \tilde{c}_i, \tilde{s}_i)).$$



Let us now return to the sequence model  $Y_n = X_n + Z_n/\sqrt{n}$  and find the limiting value of these “atomic” units as  $n \rightarrow \infty$ , and discover a simple formula for the asymptotic loss  $L_\infty(\eta|\mathbf{x})$ .

Each of these “atomic” units only depend on  $y_i$ , the  $i$ -th data singular value, and on  $c_i$  (resp.  $\tilde{c}_i$ ), the angle between the  $i$ -th left (resp. right) signal and data singular vectors. In the special case of Frobenius norm loss, we have already encountered this phenomenon (Lemma 3), where we have seen that these quantities converge to deterministic functions that depend on  $x_i$ , the  $i$ -th signal singular value alone.

For the sequence  $Y_n = X_n + Z_n/\sqrt{n}$ , recall our notation  $y_{n,i}$ ,  $\mathbf{u}_{n,i}$ ,  $\tilde{\mathbf{u}}_{n,i}$ ,  $\mathbf{v}_{n,i}$ ,  $\tilde{\mathbf{v}}_{n,i}$  from (14) and (15), and define

$$\begin{aligned} c_{n,i} &= \langle \mathbf{u}_{n,i}, \mathbf{v}_{n,i} \rangle & s_i &= \sqrt{1 - c_{n,i}^2} \\ \tilde{c}_{n,i} &= \langle \tilde{\mathbf{u}}_{n,i}, \tilde{\mathbf{v}}_{n,i} \rangle & \tilde{s}_i &= \sqrt{1 - \tilde{c}_{n,i}^2}, \end{aligned}$$

for  $i = 1, \dots, r$ . Combining Lemma 5, Lemma 1 and Lemma 2 we obtain:

**Lemma 7.** Let  $Y_n = X_n + Z_n/\sqrt{n}$  be a matrix sequence in our asymptotic framework with signal singular values  $\mathbf{x} = (x_1, \dots, x_r)$ . Assume that  $L_{2,2}$  is a regular loss and  $\eta$  is continuous at  $y(x_i)$  for some fixed  $1 \leq i \leq r$ . If  $\beta^{1/4} \leq x_i$  then

$$\lim_{n \rightarrow \infty} L_{2,2}(A(x_i), B(\eta(y_{n,i}), c_{n,i}, s_{n,i}, \tilde{c}_{n,i}, \tilde{s}_{n,i})) \stackrel{a.s.}{=} L_{2,2}(A(x_i), B(\eta(y(x_i)), x_i)),$$

where  $B(\eta, x)$  is given by (24), while if  $0 \leq x_i < \beta^{1/4}$  then

$$\begin{aligned} \lim_{n \rightarrow \infty} L_{2,2}(A(x_i), B(\eta(y_{n,i}), c_{n,i}, s_{n,i}, \tilde{c}_{n,i}, \tilde{s}_{n,i})) &\stackrel{a.s.}{=} L_{2,2}(A(x_i), \text{diag}(0, \eta(y(x_i)))) \\ &= L_{1,1}(x_i, 0) + L_{1,1}(0, \eta(y(x_i))). \end{aligned}$$

As a result, we now obtain the asymptotic loss  $L_\infty$  as a deterministic function of the nonzero population principal values  $x_1, \dots, x_r$ .

**Lemma 8. A formula for the asymptotic loss of a proper shrinker.** Assume that  $L = \{L_{m,n}\}$  is a sum- or max- decomposable family of regular and orthogonally invariant losses. Extend the definition of  $B(\eta, x)$  from (24) by setting  $B(\eta, x) = \text{diag}(0, \eta)$  for  $0 \leq x < \beta^{1/4}$ . If  $\eta : [0, \infty) \rightarrow [0, \infty)$  is a proper shrinker, then

$$L_\infty(\eta|\mathbf{x}) = \sum_{i=1}^r L_{2,2}(A(x_i), B(\eta(y(x_i)), x_i)) \quad (35)$$

if  $L$  is sum-decomposable, or

$$L_\infty(\eta|\mathbf{x}) = \max_{i=1}^r L_{2,2}(A(x_i), B(\eta(y(x_i)), x_i)) \quad (36)$$

if  $L$  is max-decomposable.

*Proof.* Consider the case  $r = 1$ . Define  $\tilde{X}_\eta(Y_n) = \sum_{i=1}^r \eta(y_{n,i}) \mathbf{v}_{n,i} \tilde{\mathbf{v}}_{n,i}'$ . Combining Lemma 7 and Lemma 5, we have

$$\lim_{n \rightarrow \infty} L_{m,n}(X, \tilde{X}_\eta(Y_n)) = L_{2,2}(A(x_1), B(\eta(y(x_1)), x_1)).$$

However,

$$\left\| \hat{X}_\eta(Y_n) - \tilde{X}_\eta(Y_n) \right\|_F^2 = \sum_{i=r+1}^{m_n} \eta(y_{n,i})^2.$$

The result now follows from Lemma 4 combined with the fact that  $L_{m_n,n}^\alpha(\cdot, \cdot)$  is Lipschitz for some  $\alpha$ .  $\square$

The final step toward the proof of Theorem 1 involves the case when the shrinker  $\eta$  is given as a special concatenation of two proper shrinkers. Even if the two part of  $\eta$  do not match, forming a discontinuity, we may still have a formula for the asymptotic loss if the loss functions match.

**Definition 8.** Assume that there exist a point  $0 < x^*$  and two shrinkers,  $\eta_1 : [0, x^*) \rightarrow [0, \infty)$  and  $\eta_2 : [x^*, \infty) \rightarrow [0, \infty)$ , such that

$$L_{2,2}\left(A(x^*), B(\eta_1(y(x^*)), x^*)\right) = L_{2,2}\left(A(x^*), B(\eta_2(y(x^*)), x^*)\right).$$

We say that the asymptotic loss functions of  $\eta_1$  and  $\eta_2$  *cross* at  $x^*$ .

**Lemma 9. A formula for the asymptotic loss of a concatenation of two proper shrinkers.** Assume that  $L = \{L_{m,n}\}$  is a sum- or max- decomposable family of regular and orthogonally invariant losses. Extend the definition of  $B(\eta, x)$  from (24) by setting  $B(\eta, x) = \text{diag}(0, \eta)$  for  $0 \leq x < \beta^{1/4}$ . Assume that there exist two shrinkers,  $\eta_1 : [0, x^*) \rightarrow [0, \infty)$  and  $\eta_2 : [x^*, \infty) \rightarrow [0, \infty)$ , whose asymptotic loss functions cross at some point  $0 < x^*$ . Define

$$\eta(x) = \begin{cases} \eta_1(x) & 0 \leq x \leq x^* \\ \eta_2(x) & x^* < x \end{cases}.$$

Then  $L_\infty(\eta|\cdot)$  exists and is given by (35) if  $L$  is sum-decomposable, or (36) if  $L$  is max-decomposable.

**Proof of Theorem 1.** Consider the shrinker  $\eta_1 \equiv 0$ . By Lemma 7,  $\eta_1$  dominates any other proper shrinker when  $0 \leq x < \beta^{1/4}$ . By assumption, there exists a point  $\beta^{1/4} \leq x_0$  such that  $\eta_1$  also dominates any proper shrinker on  $[\beta^{1/4}, x_0)$ , and such that  $\eta^{**}$  dominates any other proper shrinker on  $[x_0, \infty)$ . Finally, by assumption, the asymptotic loss functions of  $\eta_1$  and  $\eta^{**}$  cross at  $x_0$ . By Lemma 9, the concatenated shrinker  $\eta^*$  dominates any proper shrinker on  $[0, \infty)$ .  $\square$

## 6 Optimal Shrinkers for Frobenius, Operator & Nuclear Losses

Theorem 1 provides a general recipe for finding optimal singular value shrinkers. To see it in action, we turn to our three primary examples, namely, the Frobenius norm loss, the operator norm loss and the nuclear norm loss. In this section we find explicit formulas for the optimal singular value shrinkers in each of these losses.

We will need the following lemmas regarding 2-by-2 matrices (see [21]):

**Lemma 10.** The eigenvalues of any 2-by-2 matrix  $M$  with trace  $\text{trace}(M)$  and determinant  $\det(M)$  are given by

$$\lambda_{\pm}(M) = \frac{1}{2} \left( \text{trace}(M) \pm \sqrt{\text{trace}(M)^2 - 4\det(M)} \right). \quad (37)$$

*Proof.* These are the roots of the characteristic polynomial of  $M$ .  $\square$

**Lemma 11.** Let  $\Delta$  be a 2-by-2 matrix with singular values  $\sigma_+ > \sigma_- > 0$ . Define  $t = \text{trace}(\Delta\Delta') = \|\Delta\|_F^2$ ,  $d = \det(\Delta)$  and  $r^2 = t^2 - 4d^2$ . Assume that  $\Delta$  depends on a parameter  $\eta$  and let  $\dot{\sigma}_{\pm}$ ,  $\dot{t}$  and  $\dot{d}$  denote the derivative of these quantities w.r.t the parameter  $\eta$ . Then

$$r(\dot{\sigma}_+ + \dot{\sigma}_-)(\dot{\sigma}_+ - \dot{\sigma}_-) = 2(\dot{t} + 2\dot{d})(\dot{t} - 2\dot{d}).$$

*Proof.* By Lemma 10 we have  $2\sigma_{\pm}^2 = t \pm r$  and therefore

$$\sqrt{2}\dot{\sigma}_{\pm} = \frac{\dot{t} \pm \dot{r}}{2\sqrt{t \pm r}}.$$

Differentiating and expanding  $\dot{\sigma}_+ \pm \dot{\sigma}_-$  we obtain the relation

$$(\dot{\sigma}_+ + \dot{\sigma}_-) = \frac{(8d/r)(\dot{t} + 2\dot{d})(\dot{t} - 2\dot{d})}{(t^2 - r^2)(\dot{\sigma}_+ - \dot{\sigma}_-)} \quad (38)$$

and the result follows.  $\square$

**Correlary 3.** Let  $\eta, c, \tilde{c} \geq 0$  and set  $s = \sqrt{1 - c^2}$  and  $\tilde{s} = \sqrt{1 - \tilde{c}^2}$ . Define

$$\Delta = \Delta(\eta, c, \tilde{c}, s, \tilde{s}) = \begin{bmatrix} \eta c \tilde{c} - x & \eta c \tilde{s} \\ \eta \tilde{c} s & \eta s \tilde{s} \end{bmatrix}$$

Then

$$\|\Delta\|_F^2 = \eta^2 + x^2 - 2x\eta c \tilde{c} \quad (39)$$

$$\det(\Delta) = -x\eta s \tilde{s}, \quad (40)$$

and the singular values  $\sigma_+ > \sigma_-$  of  $\Delta$  are given by

$$\sigma_{\pm} = \frac{1}{\sqrt{2}} \sqrt{\|\Delta\|_F^2 \pm \sqrt{\|\Delta\|_F^4 - 4\det(\Delta)^2}}. \quad (41)$$

## 6.1 Frobenius norm loss

Theorem 1 allows us to rediscover the optimal shrinker for Frobenius norm loss, which we derived from first principles in Section 4. To this end, observe that by (39) we have

$$L_{2,2}^{fro} \left( \eta \begin{bmatrix} c\tilde{c} & c\tilde{s} \\ \tilde{c}s & s\tilde{s} \end{bmatrix}, \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right) = \|\Delta\|_F^2 = \eta^2 + x^2 - 2x\eta c \tilde{c}. \quad (42)$$

To find the optimal shrinker, we solve  $\partial \|\Delta\|_F^2 / \partial \eta = 0$  for  $\eta$  and use the fact that  $c^2 \tilde{c}^2 + c^2 \tilde{s}^2 + s^2 \tilde{c}^2 + s^2 \tilde{s}^2 = 1$ . We find that the minimizer of  $\|\Delta\|_F^2$  is  $\eta^{**}(x) = x c \tilde{c}$ . Defining  $\eta^{**}(x) = x c(x) \tilde{c}(x)$  for  $x \geq \beta^{1/4}$ , we find that the asymptotic loss of  $\eta^{**}(x)$  and of  $\eta \equiv 0$  cross at  $x_0 = \beta^{1/4}$ . Simplifying (28), where  $x(y)$  is given by (5), we find that  $\eta^*(y)$  is given by (4), and is in fact a proper shrinker. By Theorem 1, this is an optimal (proper) shrinker.

## 6.2 Operator norm loss

By (41),

$$L_{2,2}^{op} \left( \eta \begin{bmatrix} c\tilde{c} & c\tilde{s} \\ \tilde{c}s & s\tilde{s} \end{bmatrix}, \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right) = \|\Delta\|_{op} = \sigma_+. \quad (43)$$

To find the optimal shrinker, we solve  $\partial \|\Delta\|_{op} / \partial \eta = 0$  for  $\eta$  on  $\beta^{1/4} \leq x$ . We find that the minimizer of  $\|\Delta\|_{op}$  is at  $\eta^{**}(x) = x$ . By (41), the asymptotic loss of  $\eta^{**}$  is given by  $x\sqrt{1 - c(x)\tilde{c}(x) + |c(x) - \tilde{c}(x)|}$ , so that the asymptotic loss of  $\eta^{**}(x)$  and of  $\eta \equiv 0$  cross at  $x_0 = \beta^{1/4}$ . Simplifying (28), we recover that the shrinker  $\eta^*(y)$  in (6). Observe that although this shrinker is discontinuous at  $y(\beta^{1/4}) = 1 + \sqrt{\beta}$ , its asymptotic loss  $L_\infty(\eta^*|\cdot)$  exists, and, by Theorem 1, dominates any proper shrinker.

**Remark.** The optimal shrinker for operator norm loss  $\eta^*(y) = x(y)$  simply shrinks the data singular value back to the "original" location of its corresponding signal singular value. A similar phenomenon has been observed in eigenvalue shrinkage for covariance estimation, see [21].

## 6.3 Nuclear norm loss

Again by (41)

$$L_{2,2}^{nuc} \left( \eta \begin{bmatrix} c\tilde{c} & c\tilde{s} \\ \tilde{c}s & s\tilde{s} \end{bmatrix}, \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right) = \|\Delta\|_* = \sigma_+ + \sigma_-. \quad (44)$$

To find the optimal shrinker, assume first that  $x$  is such that  $c(x)\tilde{c}(x) \geq s(x)\tilde{s}(x)$ . By Lemma 11, with

$$t = \eta^2 + x^2 - 2x\eta c\tilde{c} \quad \text{and} \quad d = -x\eta s\tilde{s},$$

we find that only zero of  $\partial(\sigma_+ + \sigma_-)/\partial \eta$  occurs when  $\partial t/\partial \eta - \partial d/\partial \eta = 0$ , namely at

$$\eta^{**}(x) = x \left( c(x)\tilde{c}(x) - s(x)\tilde{s}(x) \right).$$

Direct calculation using (41), (40) and (39) shows that the square of the asymptotic loss of  $\eta^{**}$  is simply  $x^2 + (\eta^{**}(x))^2 - 2x\eta^{**}(x) \left( c(x)\tilde{c}(x) - s(x)\tilde{s}(x) \right)$ , so that the asymptotic loss of  $\eta^{**}(x)$  and of  $\eta \equiv 0$  cross at the unique  $x_0$  satisfying  $c(x_0)\tilde{c}(x_0) = s(x_0)\tilde{s}(x_0)$ . Substituting  $c = c(x)$  from (12),  $\tilde{c} = \tilde{c}(x)$  from (13) and also  $s = s(x) = \sqrt{1 - c(x)^2}$  and  $\tilde{s} = \tilde{s}(x) = \sqrt{1 - \tilde{c}(x)^2}$ , we find

$$\eta^*(y) = \left( \frac{x^4 - \beta}{x^2 y} + \frac{\sqrt{\beta}}{x} \right)_+,$$

recovering the optimal shrinker (6).

## 7 Extensions

Our main results have been formulated and calibrated specifically for the model  $Y = X + Z/\sqrt{n} \in M_{m \times n}$ , where the distribution of the noise matrix  $Z$  is orthogonally invariant. In this section we extend our main results to include the model  $Y = X + \sigma Z \in M_{m \times n}$ , and consider:

1. The setting where  $\sigma$  is either known but does not necessarily equal  $1/\sqrt{n}$ , or is altogether unknown.
2. The setting where the noise matrix  $Z$  has i.i.d entries, but its distribution is not necessarily orthogonally invariant.

The results below follow [13].

### 7.1 Unknown Noise level

Consider an asymptotic framework slightly more general than the one in Section 2.2, in which  $Y_n = X_n + (\sigma/\sqrt{n})Z_n$ , with  $X_n$  and  $Z_n$  as defined there. In this section we keep the loss family  $L$  and the asymptotic aspect ratio  $\beta$  fixed and implicit. We extend Definition 1 and write

$$L_\infty(\eta|\mathbf{x}, \sigma) \stackrel{a.s.}{=} \lim_{n \rightarrow \infty} L_{m_n, n} \left( X_n, \hat{X}_\eta(X_n + \frac{\sigma}{\sqrt{n}}Z_n) \right).$$

When the noise level  $\sigma$  is known, Eq. (8) allows us to re-calibrate any nonlinearity  $\eta$ , originally calibrated for noise level  $1/\sqrt{n}$ , to a different noise level. For a nonlinearity  $\eta : [0, \infty) \rightarrow [0, \infty)$ , write

$$\eta_c(y) = c \cdot \eta(y/c).$$

We clearly have:

**Lemma 12.** If  $\eta^*$  is an optimal shrinker for  $Y_n = X_n + Z_n/\sqrt{n}$ , namely,

$$L_\infty(\eta^*|\mathbf{x}) \leq L_\infty(\eta|\mathbf{x})$$

for any  $r \geq 1$ , any  $\mathbf{x} \in \mathbb{R}^r$  and any continuous nonlinearity  $\eta$ , and  $\sigma > 0$ , then  $\eta_\sigma^*$  is an optimal shrinker for  $Y_n = X_n + (\sigma/\sqrt{n})Z_n$ , namely

$$L_\infty(\eta_\sigma^*|\mathbf{x}, \sigma) \leq L_\infty(\eta|\mathbf{x}, \sigma)$$

for any  $r \geq 1$ , any  $\mathbf{x} \in \mathbb{R}^r$  and any continuous nonlinearity  $\eta$ .

When the noise level  $\sigma$  is unknown, we are required to estimate it. See [14, 26] and references therein for existing literature on this estimation problem. The method below has been proposed in [13].

Consider the following robust estimator for the parameter  $\sigma$  in the model  $Y = X + \sigma Z$ :

$$\hat{\sigma}(Y) = \frac{y_{med}}{\sqrt{n} \cdot \mu_\beta}, \quad (45)$$

where  $y_{med}$  is a median singular value of  $Y$  and  $\mu_\beta$  is the median of the Marcenko-Pastur distribution, namely, the unique solution in  $\beta_- \leq x \leq \beta_+$  to the equation

$$\int_{\beta_-}^x \frac{\sqrt{(\beta_+ - t)(t - \beta_-)}}{2\pi t} dt = \frac{1}{2},$$

where  $\beta_\pm = 1 \pm \sqrt{\beta}$ . Note that the median  $\mu_\beta$  is not available analytically but can easily be obtained by numerical quadrature.

**Lemma 13.** Let  $\sigma > 0$ . For the sequence  $Y_n = X_n + (\sigma/\sqrt{n})Z_n$  in our asymptotic framework,

$$\lim_{n \rightarrow \infty} \frac{\hat{\sigma}(Y_n)}{1/\sqrt{n}} \stackrel{a.s.}{=} \sigma.$$

**Correlary 4.** Let  $Y_n = X_n + (\sigma/\sqrt{n})Z_n$  be a sequence in our asymptotic framework and let  $\eta^*$  be an optimal shrinker calibrated for  $Y_n = X_n + Z_n/\sqrt{n}$ . Then the random sequence of shrinkers  $\eta_{\hat{\sigma}(Y_n)}^*$  converges to the optimal shrinker  $\eta_\sigma^*$ :

$$\lim_{n \rightarrow \infty} \eta_{\hat{\sigma}(Y_n)}(y) \stackrel{a.s.}{=} \eta_\sigma(y), \quad y > 0.$$

Consequently,  $\eta_{\hat{\sigma}(Y_n)}^*$  asymptotically achieves optimal performance:

$$\lim_{n \rightarrow \infty} L_{m,n} \left( X_n, \hat{X}_{\eta_{\hat{\sigma}(Y_n)}^*} \left( X_n + \frac{\sigma}{\sqrt{n}} Z_n \right) \right) = L_\infty(\eta_\sigma^* | \mathbf{x}, \sigma).$$

In practice, for denoising a matrix  $Y \in M_{m \times n}$ , assumed to satisfy  $Y = X + \sigma Z$ , where  $X$  is low-rank and  $Z$  has i.i.d entries, we have the following approximately optimal singular value shrinkage estimator:

$$\hat{X}(Y) = \sqrt{n}\sigma \hat{X}_{\eta^*}(Y/(\sqrt{n}\sigma)) \quad (46)$$

when  $\sigma$  is known, and

$$\hat{X}(Y) = \sqrt{n}\hat{\sigma}(Y) \cdot \hat{X}_{\eta^*}(Y/(\sqrt{n}\hat{\sigma}(Y))) \quad (47)$$

when  $\sigma$  is unknown. Here,  $\eta^*$  is an optimal shrinker with respect to desired loss family  $L$  in the natural scaling.

## 7.2 General white noise

Our results were formally stated for the sequence of models of the form  $Y = X + \sigma Z$ , where  $X$  is a non-random matrix to be estimated, and the entries of  $Z$  are i.i.d samples from a distribution that is orthogonally invariant (in the sense that the matrix  $Z$  follows the same distribution as  $AZB$ , for any orthogonal  $A \in M_{m,m}$  and  $B \in M_{n,n}$ ). While Gaussian noise is orthogonally invariant, many common distributions, which one could consider to model white observation noise, are not.

The singular values of a signal matrix  $X$  are a very widely used measure of the complexity of  $X$ . In particular, they capture its rank. One attractive feature of the framework we adopt is that the loss  $L_{m,n}(X, \hat{X})$  only depends on the signal matrix  $X$  through its nonzero singular values  $\mathbf{x}$ . This allows the loss to be directly related

to the complexity of the signal  $X$ . If the distribution of  $Z$  is not orthogonally invariant, the loss no longer has this property. This point is discussed extensively in [14].

In general white noise, which is not necessarily orthogonally invariant, one can still allow the loss to depend on  $X$  only through its singular values by placing a prior distribution on  $X$  and shifting to a model where it is a random, instead of a fixed, matrix. Specifically, consider an alternative asymptotic framework to the one in Section 2.2, in which the sequence denoising problems  $Y_n = X_n + Z_n/\sqrt{n}$  satisfies the following assumptions:

1. *General white noise:* The entries of  $Z_n$  are i.i.d samples from a distribution with zero mean, unit variance and finite fourth moment.
2. *Fixed signal column span and uniformly distributed signal singular vectors:* Let the rank  $r > 0$  be fixed and choose a vector  $\mathbf{x} \in \mathbb{R}^r$  with coordinates  $\mathbf{x} = (x_1, \dots, x_r)$ . Assume that for all  $n$ ,

$$X_n = U_n \text{diag}(x_1, \dots, x_r, 0, \dots, 0) V_n' \quad (48)$$

is a singular value decomposition of  $X_n$ , where  $U_n$  and  $V_n$  are uniformly distributed random orthogonal matrices. Formally,  $U_n$  and  $V_n$  are sampled from the Haar distribution on the  $m$ -by- $m$  and  $n$ -by- $n$  orthogonal group, respectively.

3. *Asymptotic aspect ratio  $\beta$ :* The sequence  $m_n$  is such that  $m_n/n \rightarrow \beta$ .

The second assumption above implies that  $X_n$  is a “generic” choice of matrix with nonzero singular values  $\mathbf{x}$ , or equivalently, a generic choice of coordinate systems in which the linear operator corresponding to  $X$  is expressed.

The results of [16], which we have used, hold in this case as well. It follows that Lemma 1 and Lemma 2, and consequently all our main results, hold under this alternative framework. In short, in general white noise, all our results hold if one is willing to only specify the signal singular values, rather than the signal matrix, and consider a “generic” signal matrix with these singular values.

## 8 Conclusion

We have presented a general method to show that optimal shrinkers exist for a variety of loss functions, and to find them. While in each of the three example loss functions we discussed there exists a simple, explicit formula for the optimal shrinker, that is not always the case. The minimization problem (26) may not admit a simple analytic solution, but it is still extremely easy to solve numerically. In these cases, implementing our method would involve numerical tabulation of the optimal shrinker.

Note that our general method, Theorem 1, finds a shrinker that is asymptotically unique admissible, or optimal, among *proper* shrinkers (in the sense of Definition 4), rather than among all continuous shrinkers. This is an artifact of our general approach, which handles a large variety of loss families in a single theorem. Shrinkers that are continuous but not regular are strange objects, and indeed optimal shrinkers formally derived using our method may well be optimal among all continuous shrinkers, although proving this seems to require



delicate arguments tailored to individual losses. For example, in the introductory Section 4 we have proved that the shrinker (4), which was also derived using our general framework, is indeed optimal among all continuous shrinkers.

Finding an explicit form of the optimal shrinkers for other matrix loss families, such as other Schatten- $p$  losses (see [20] and references within), and a generalization of our method to include Ky-Fan norms, both remain interesting problems for further study.

## Acknowledgements

This work was partially supported by NSF DMS-0906812 (ARRA). We thank Iain Johnstone for helpful comments. We also thank Amit Singer and Boaz Nadler for discussions stimulating this work, and Santiago Velasco-Forero for pointing out an error in an earlier version of the manuscript. MG was partially supported by a William R. and Sara Hart Kimball Stanford Graduate Fellowship.

## References

- [1] Gene H. Golub and W Kahan. Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial & Applied Mathematics: Series B*, 2(2):205–224, 1965.
- [2] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 1936.
- [3] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, August 2000.
- [4] RB Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.
- [5] DA Jackson. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 1993.
- [6] T D Lagerlund, F W Sharbrough, and N E Busacker. Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 14(1):73–82, January 1997.
- [7] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy a Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–9, August 2006.
- [8] O Edfors and M Sandell. OFDM channel estimation by singular value decomposition. *IEEE Transactions on Communications*, 46(7):931–939, 1998.
- [9] David L. Donoho and Iain M. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81(3):425–455, 1994.

- [10] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.
- [11] David L. Donoho. De-Noising by Soft-Thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, May 1995.
- [12] David L. Donoho and Iain M. Johnstone. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B*, 57(2):301–369, 1995.
- [13] Matan Gavish and David L. Donoho. The Optimal Hard Threshold for Singular Values is  $4/\sqrt{3}$ . *IEEE Transactions on Information Theory*, to appear (*arXiv:1305.5870*), 2013.
- [14] Andrey A. Shabalin and Andrew B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, July 2013.
- [15] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.
- [16] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, October 2012.
- [17] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, December 2008.
- [18] Clifford Lam and Jianqing Fan. Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation. *Annals of statistics*, 37(6B):4254–4278, January 2009.
- [19] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, August 2010.
- [20] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.
- [21] David L. Donoho, Matan Gavish, and Iain M. Johnstone. Optimal shrinkage of eigenvalues in the Spiked Covariance Model. *arXiv:1311.0851*, 2013.
- [22] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge university press, Cambridge, 2010.
- [23] Debashis Paul. Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. *Statistica Sinica*, 17(4):1617–1642, 2007.
- [24] R. Brent Dozier and Jack W. Silverstein. On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, 98(4):678–694, April 2007.
- [25] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

- 
- [26] Shira Kritchman and Boaz Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *Signal Processing, IEEE Transactions*, 57(10):3930–3941, 2009.